Volume 6 Number 7 1979 Nucleic Acids Research

A strategy of DNA sequencing employing computer programs

R.Staden

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 23 March 1979

ABSTRACT

With modern fast sequencing techniques[1,2] and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps. This paper describes computer programs that can be used to order both sequence gel readings and clones. A method of coding for uncertainties in gel readings is described. These programs are available on request.

INTRODUCTION

It became clear during the sequencing of bacteriophage ØX174 DNA[3] that it was necessary to use computers to handle and analyse the data. Later, while the very similar DNA sequence of bacteriophage G4 [4] was being determined, the computer was used to compare and align the G4 sequence with that of ØX174. Further advances in DNA sequencing methods[5,6] and cloning techniques have been made and work is in progress in a number of laboratories on sequences many times the length of ØX174.

The continuing rapid fall in the cost of computer components is making it possible for most DNA sequencing laboratories to have their own small computer. The fact that DNA sequencing is now a fast procedure, and the availability of computers gives the possibility of more efficient overall strategies for sequence determination. Outlined below is one such strategy which takes into account cloning technology, the speed of DNA sequencing and the ability of computers to handle and compare data. This is followed by a description of a method of coding for uncertainties in gel readings. The rest of the paper contains descriptions of easily used computer programs that facilitate this overall sequencing strategy or variations on it.

The Strategy

The whole of the DNA to be sequenced is shotgunned into a suitable vector and cloned. Ideally the cloned fragments would be of at least 200

bases in length. The clones are then sequenced and the computer used to collate the data. Collation involves searching for overlaps in the data.

If the 5' end of the sequence from one gel reading is the same as the 3' end of the sequence from another the data is said to overlap. If the overlap is of sufficient length to distinguish it from being a repeat in the sequence the two sequences must be contiguous. The data from the two gel readings can then be joined to form one longer continuous sequence.

To facilitate the search for overlaps all sequences derived from previous gel readings are stored in one master file. All new gel readings and their complements are compared with the sequences in the master file. Any sequences involved in overlaps are joined and new data is added into the Master file. Eventually the sequence of each clone will be finished and its contributing gel readings collated and joined. The same procedure also orders the individual clones to produce the finished sequence.

Earlier papers[7,8] described basic sequence data handling and analysis programs designed for use by those unfamiliar with computers. The programs described here have all the design features of these earlier programs and are used on the same small computer but have been extended so that they can analyse sequences of up to 18000 nucleotides. They are in daily use as an integral part of a large sequencing project in this laboratory.

Three programs are described:

1) The program OVRLAP searches for overlaps between sequences from different gel readings or clones; 2) XMATCH is a program which allows the operator to scrutinise overlapping sequences, edit them and join them together; 3) FILINS is a general program for the manipulation of sequence files.

The uncertainty codes

The use of a simple 5 letter code A,C,G,T, - for recording nucleotide sequences is inflexible as it does not permit doubts to be recorded. One of the problems of using computers is that once the data has been filed away in the machine its quality is often forgotten: any data in the machine is taken to be correct although originally doubts may have existed. The uncertainty codes allow these doubts to be recorded until such time as new experimental data allow them to be removed. In order to get the most from the gels and yet maintain a high level of accuracy in the data we have introduced an 18 character code. The 18 extra characters allow us to code for, and hence keep track of all types of uncertainty in the data. The code is shown in Fig. 1.

Fig 1.

Symbols for uncertain nucleotide sequences

Symbol	Meaning
1	Probably C
2	" T
3	" A
4	" G
D	Definitely C possibly CC
V	" T " TT
B	" A " AA
H	" G " GG
K	Definitely C possibly CX
L	" T " TX
M	" A " AX
N	" G " GX
R	A or G
Y	C or T
5	A or C
6	G or T
7	A or T
8	G or C
-	A or C or G or T

Description of the programs

In the examples shown all operator input is underlined; all other printing is done by the programs.

1) OVRLAP

This program is used for comparing sequences derived from new gel readings with a master file containing the data from all previous gel readings in order to search for overlaps. A minimum length of overlap is specified by the operator and any overlaps of at least this length are reported by the program. The program also calculates and compares the complement of the new gel reading with the master file.

An example is shown in Fig. 2. The operator has supplied the names of the files containing the new gel reading and the master file. He has then asked the program to compare the whole of each of these by not defining a restricted region from the master file or string from the new gel reading. He has asked the program only to report overlaps of at least 10 characters. The program has found 3 overlaps on the input strand and none on its complement. Each overlap is described by a column of numbers giving the number of matching characters, and the positions of the first character in the overlap, in each sequence.

2) XMATCH

When an overlap is found between two gel readings further operations need to be performed.

Fig 2.

```
RU OVRLAP

        PLEASE TYPE NAME OF FILE 1
MASTER.RS2

        PLEASE TYPE NAME OF FILE 2
LOUT7

        REGION
FIRST SEQ NO = __

LAST SEQ NO = __

        STRING
FIRST SEQ NO = _

LAST SEQ NO = _

   MIN =  10

   MATCHES FOUND =      3

      26    16    32
    5477  5394  5422
     162    77   105

   MATCHES FOUND =      0
```

a) The overlap must be scrutinized to determine if it is from two contiguous sequences or simply a repeat in the sequence;

b) if necessary either of the two sequences may need to be edited to optimise the overlap;

c) a check must be made that any uncertainty codes that are superposed by the overlap are in agreement;

d) then we must produce, for the overlapping region, a sequence that is in agreement with both gel readings and contains the best from each. (This we term the 'best' sequence).

e) Finally we must join the two gel readings to make one longer sequence, the overlap being replaced by the 'best' sequence.

XMATCH is a program for performing all these tasks. It is most efficiently operated from a V.D.U. but any interactive terminal will suffice.

The program performs these tasks using five commands:

I. Insert allows insertion of characters into either sequence.

II. Delete allows deletion of characters from either sequence.

III. Next allows display of a different pair of sequence sections.

IV. Best instructs the program to produce the 'best' sequence from the two overlapping sections. Disagreements or mismatches are shown by asterisks.

V. Save instructs the program to join the two gel readings together and save them on a disk file.

After each manipulation the program immediately displays the result on the V.D.U. screen or keyboard. After the commands Next, Insert, Delete one sequence is displayed above the other with identities designated by asterisks. Following the command Best one sequence is displayed above the other with the 'best' sequence between them; in this case mismatches are marked with asterisks. An example of the use of this program is shown in Fig. 3 where it is used to operate on the overlaps found in Fig. 2. Sequence number 1 is the master file and sequence 2 is the new gel reading.

The operator supplies the names of the two files containing the overlap and the sequence positions of the first characters in the overlap. The program displays these two sections of sequence one above the other to the end of the new gel reading. Identical characters are marked by asterisks. The operator then selects the Next command to examine the region immediately to the left of the overlap. When this is displayed it can be seen that the overlap has started at the beginning of a gel reading in the master file. (Individual gel readings in the master file are separated by titles, enclosed

Fig 3.

```
$RU XMATCH

        PLEASE TYPE NAME OF FILE 1
MASTER.RS2

        PLEASE TYPE NAME OF FILE 2
LOUT7

 FIRST SEQ NO =5394

 LAST SEQ NO = _

 FIRST SEQ NO =77

 LAST SEQ NO = _


     5394
        GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
        ********** ****** **  ******* ** ********** ********** **********
        GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
       77
     5454
        HATT1THCCG 31GGGCTA1G -GYATT1CTG CGTTAA1CTG TTCCATCGTG G
           *          *       *          *         *        *
        GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
      137
 COMMAND(D,I,N,B,S) =N

 FIRST SEQ NO =5390

 LAST SEQ NO = _

 FIRST SEQ NO =73

 LAST SEQ NO =80


     5390
        --->GTAA
            ****
        GGCGGTAA
       73
 COMMAND(D,I,N,B,S) =N

 FIRST SEQ NO =5394

 LAST SEQ NO = _

 FIRST SEQ NO =77

 LAST SEQ NO = _
```

in <...> characters.) The overlap is therefore of the form

```
MMMMMMMM
NNNNNNNN
```

where M characters represent the master file gel reading and N characters

represent the new gel reading.

The operator then selects the Next command to define the pieces of sequence he wishes to manipulate to optimise the match. Inspection of this region shows an H character at position 5454 in sequence 1: this codes for

Fig 3 cont.

```
    5394
       GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
       ********** ****** **  ******* ** ********** ********** **********
       GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         77
    5454
       HATT1THCCG 31GGGCTA1G -GYATT1CTG CGTTAA1CTG TTCCATCGTG G
          *          *       *          *         *        *
       GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
      137
COMMAND(D,I,N,B,S) =I

SEQNCE NUMBER(1,2) =1

POSITION =5454

CHARACTERS =G@


    5394
       GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
       ********** ****** **  ******* ** ********** ********** **********
       GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         77
    5454
       GHATT1THCC G31GGGCTA1 G-GYATT1CT GCGTTAA1CT GTTCCATCGT GG
       * *** *  *     **           *        * *      * *
       GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
      137
COMMAND(D,I,N,B,S) =I

SEQNCE NUMBER(1,2) =1

POSITION =5461

CHARACTERS =G@


    5394
       GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
       ********** ****** **  ******* ** ********** ********** **********
       GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         77
    5454
       GHATT1TGHC CG31GGGCTA 1G-GYATT1C TGCGTTAA1C TGTTCCATCG TGG
       * *** ** * **  *** **  * * ***** ********** ********** *
       GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
      137
COMMAND(D,I,N,B,S) =N

FIRST SEQ NO =5394

LAST SEQ NO = __

FIRST SEQ NO =77

LAST SEQ NO = __
```

G or GG. Putting in an extra G character would extend the match and so the operator selects the Insert command. The Insert routine asks the operator to define the sequence he wishes to operate on, where in the sequence, and then to type in the characters to insert. [The operator may type in any number of characters on any number of lines and in order to tell the program

Fig 3 cont.

```
      5394
         GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         ********** ****** **  ******* ** ********** ********** **********
         GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
        77
      5454
         GHATT1TGHC CG31GGGCTA 1G-GYATT1C TGCGTTAA1C TGTTCCATCG T
         * *** ** * **  *** **  * * ***** ********** **********  *
         GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
       137
   COMMAND(D,I,N,B,S) =B

      5394
         GTAACGGATG CTTCTT1CC4 GCA1CAT4CA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
         GTAACGGATG CTTCTTCCCG GCA1CATGCA ACAAACTGCC CGGGTGATGG CAGAAATGGT
        77
      5454
         GHATT1TGHC CG31GGGCTA 1G-GYATT1C TGCGTTAA1C TGTTCCATCG T
         GGATTCTGGC CGACGGGCTA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
         GGATTCTGGC CGACGGG1TA CGCGCATT1C TGCGTTAA1C TGTTCCATCG T
       137
   COMMAND(D,I,N,B,S) =S


    LEFT HAND END
   SEQNCE NUMBER(1,2) =2

  FIRST SEQ NO =_

  LAST SEQ NO = _


     RIGHT HAND END
  FIRST SEQ NO = _

  LAST SEQ NO = _


          PLEASE TYPE NAME OF FILE 3
 MLOUT7,1

                  10          20          30          40          50          60
          GATAACGCTC TTGGCAAATT TACAGTGTCA GATACGATAG CAGATAGATA CCTTTTAGAA

                  70          80          90         100         110         120
          GTTTGGCAGT CCGGCGGTAA CGGATGCTTC TTCCCGGCA1 CATGCAACAA ACTGCCCGGG

                 130         140         150         160         170         180
          TGATGGCAGA AATGGTGGAT TCTGGCCGAC GGGCTACGCG CATT1CTGCG TTAA1CTGTT

                 190         200         210         220         230         240
          CCATCGTGGT G3TTCCCGTT TTCCCGAAAR GCCAGAACCC ACTGGCGA1G GATTT4TTTC

                 250         260         270         280         290         300
          A1T214T212 GG2CA1GG22 AGCCAGGTT1 GGC1GGGAAA
```

he has finished typing he types the special character @.] The program then displays the result of this insertion. Another H is at position 5461 in sequence 1 and it is clear that an extra G at this point will extend the overlap so the operator again has selected the Insert command and put in the G. [Note: it is often necessary to refer back to the experimental data to make these editing decisions.] Now the operator wants to know whether any disagreements exist in the uncertainty codes. He therefore selects the Next command to define the overlap and then the Best command. The result is displayed and the best sequence is consequently written between the two gel readings. There are no disagreements (these would have been denoted by asterisks) and so now the operator selects the Save command to join the two sequences together.

The program then asks which is the lefthand sequence and gives the operator the opportunity to define a restricted part of it to add to the best sequence. In this case the operator does not define a restricted region for either the left or right flanking sequences and so the program takes the whole of each gel reading and joins the two together with the best sequence between. The program then requests that the operator gives the new sequence a file name and writes the file. It then prints a copy of the contents of the file which is shown in this figure with best sequence i.e. the overlap, overlined.

3) FILINS

This program is used to produce a final ordered sequence from a number of files containing overlapped gel readings. Alternatively it can be employed to produce the final ordered sequence for a whole genome from a number of sequenced clones. The program allows the operator to extract as many regions from as many files as he wishes and arrange them in any order. This program greatly simplifies this process.

Summary

A magnetic tape containing the FORTRAN of these and several other new programs is available on request. Each tape will be accompanied by notes on the construction, and use of each program.

ACKNOWLEDGEMENTS

I would like to thank B.G. Barrell for discussions and F. Sanger and J. Walker for advice during the preparation of this article.

REFERENCES

1 Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. 94, 441-448.
2 Maxam, A.M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
3 Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A. III, Slocombe, P.M. and Smith, M. (1977) Nature 265, 687-695.
4 Godson, G.N., Barrell, B.G., Staden, R. and Fiddes, J.C. (1978) Nature 276, 236-247.
5 Sanger, F. and Coulson, A.R. (1978) FEBS Letters 87, 107-110.
6 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
7 Staden, R. (1977) Nucleic Acids Research 4, 4037-4051.
8 Staden, R. (1978) Nucleic Acids Research 5, 1013-1015.